Three Ways to Improve the Validity of Statistical Results

Alex Harris, Ph.D.

Program Evaluation and Resource Center

Center for Health Care Evaluation

VA Palo Alto HCS

HSR&D Annual Meeting, February, 16th, 2006

Views Expressed are the Presenter’s not of the Department of Veterans Affairs

Workshop Goals

- To describe three common practices in data analysis that can lead to misleading results
- To introduce better alternatives
- Present the material so it is useful and comprehensible to a wide audience

Common Threats to the Validity of Statistical Results

- Model-Data Incongruence
  - Failure to check if model assumptions are met
    - Linear models used with non-linear relationships
    - Non-normally distributed errors
    - Non-independent errors
- Use of outdated or flawed methods
  - Advances in theory, simulation studies that lead to the emergence of better strategies
    - Repeated measures ANOVA, pair-wise deletion, some automatic variable selection techniques

Workshop Plan

- Checking the assumptions of linear models and various fixes to problems
- Discuss problems of using models that assume independence of errors when dependencies exist.
- Survey good and bad methods to handle missing data.

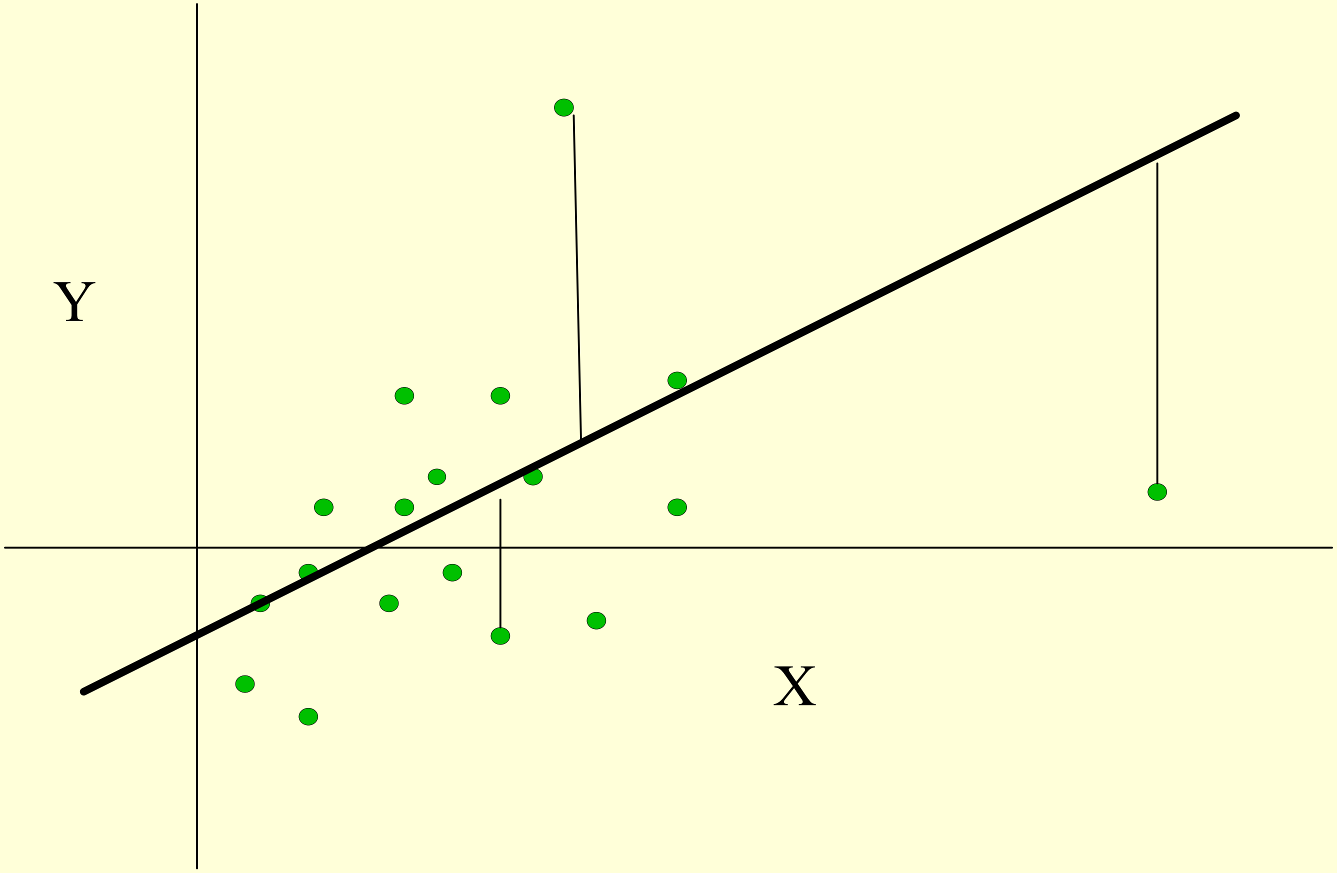
Punch Line 1

- Know the assumptions of the models you use.
- Know how robust they are to violations of assumptions?
- Know how to detect the degree to which assumptions are met.
- Learn a few strategies for addressing problematic violations of assumptions.

Assumptions of Linear Regression

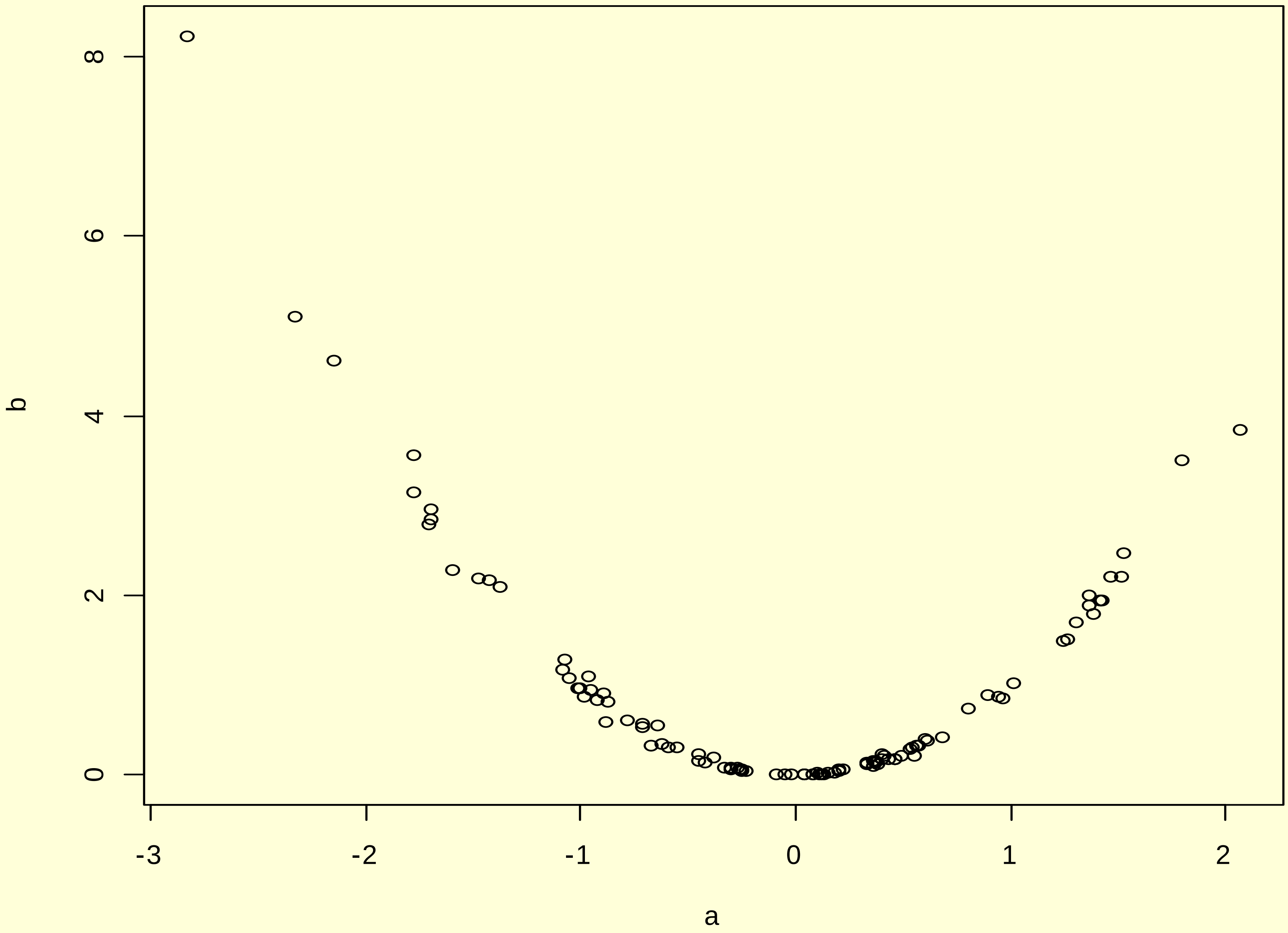
1. **Linearity** of the relationship between dependent and independent variables
2. **Independence** of the errors
3. **Homoscedasticity** (constant variance) of the errors
4. **Normality** of the error distribution.

“Error” and “Residual” refer to the same thing



Violation of the Linearity Assumption

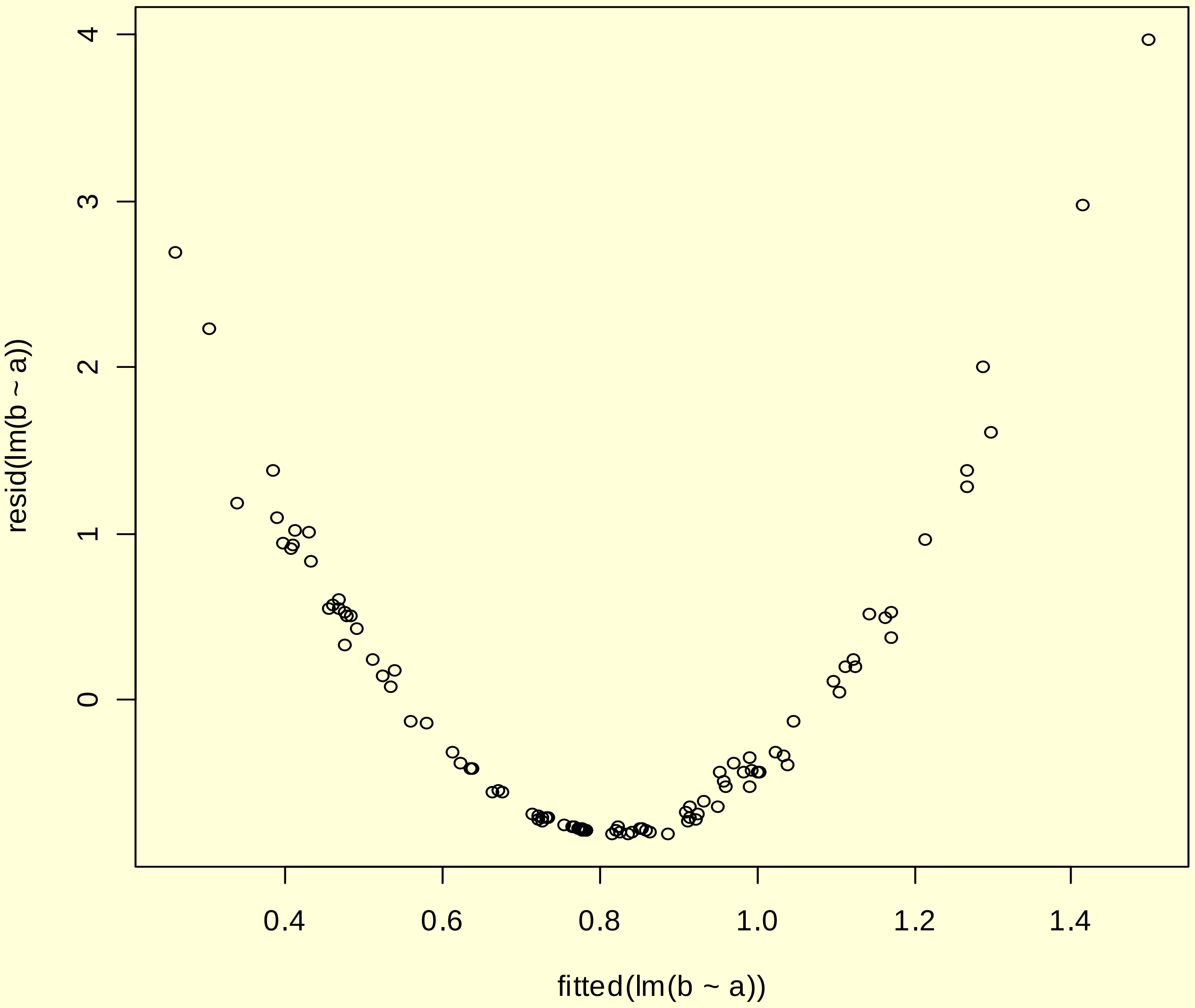
- Leads to incorrect predictions
- Leads you to conclude predictor is less associated with an outcome than it is


b
0
2
4
6
8
-3
-2
-1
0
1
2
a

```
lm(formula = b ~ a)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9317     0.1273   7.317 7.04e-11 ***
a            -0.4302     0.1298  -3.314  0.00129 **

Adjusted R-squared: 0.0916
```

How to Detect Violation in Linearity

- Look at plots
  - *Observed versus observed*
  - *Residuals versus predicted values*



resid(lm(b ~ a))
4
3
2
1
0
0.4
0.6
0.8
1.0
1.2
1.4
fitted(lm(b ~ a))

How to Fix

- Stay with the linear regression framework and
  - Add a regressor which is a non-linear function of one of the predictors
  - Apply a *nonlinear transformation* to the the dependent and/or independent les (Rule of the Bulge)

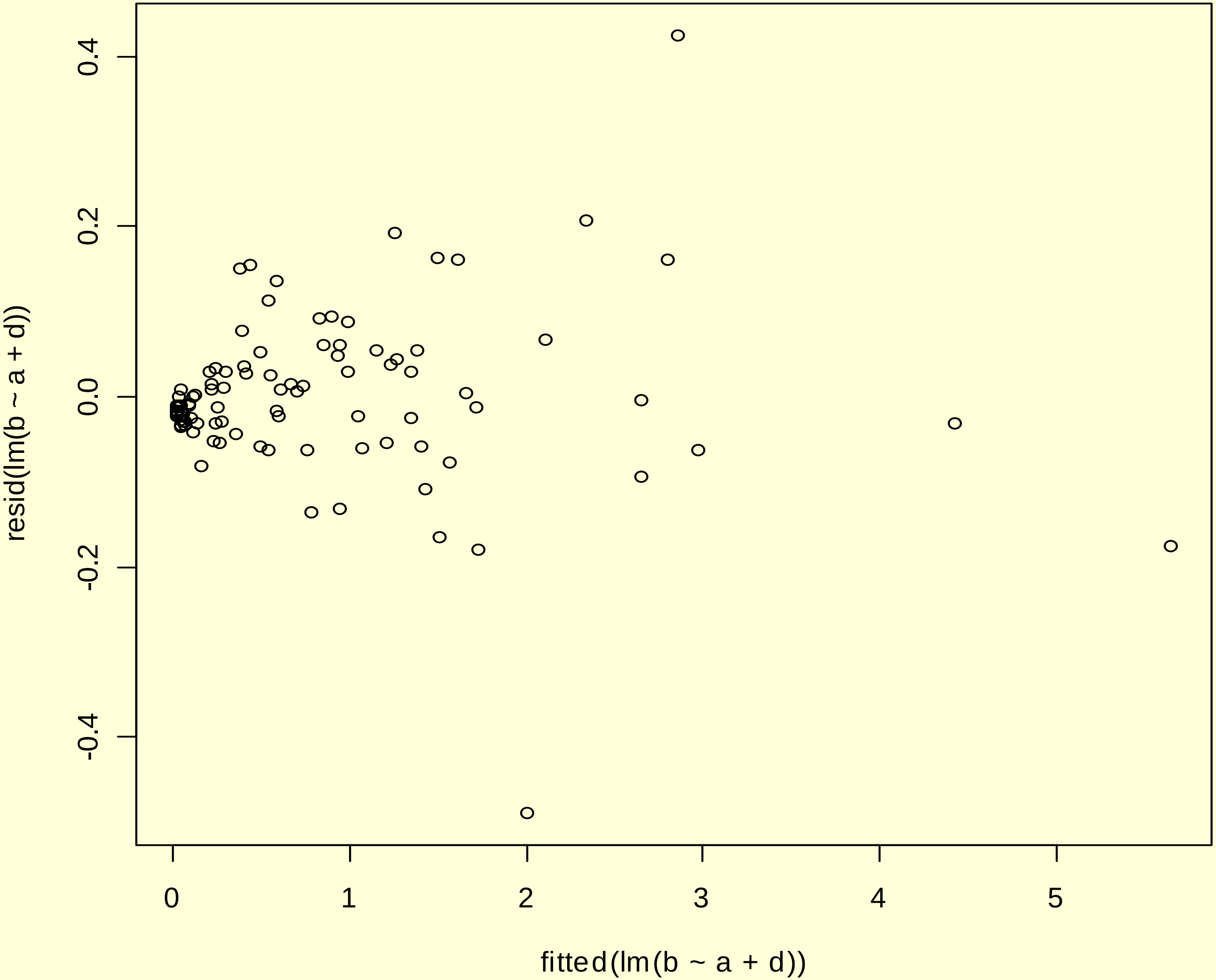
    ge)

  del

  gy

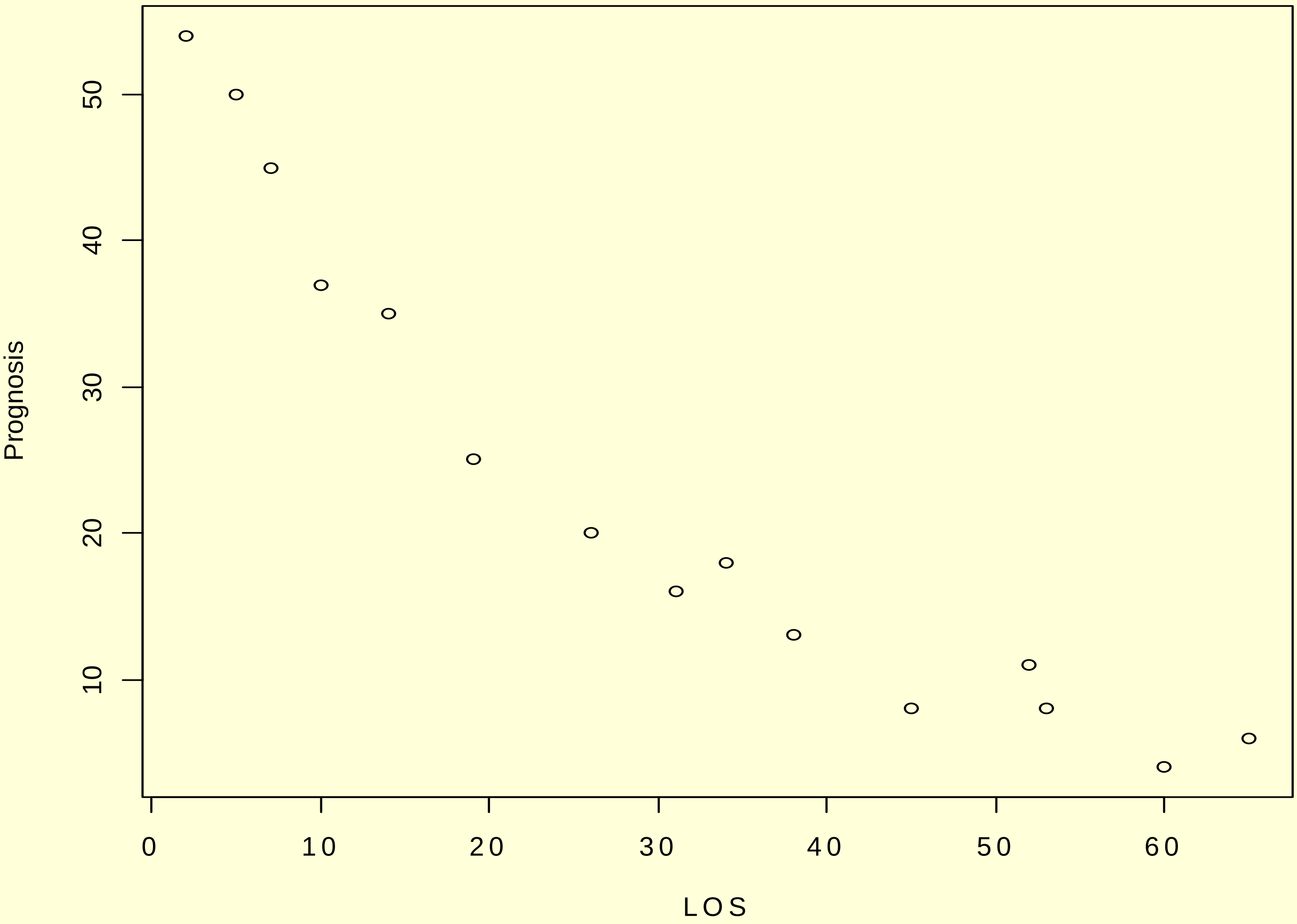
  y

```
d = a^2
lm(formula = b ~ a + d)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.020642   0.012686   1.627    0.107
a           0.003107   0.011592   0.268    0.789
d           0.991819   0.010558  93.944   <2e-16 ***

Adjusted R-squared: 0.9898
```

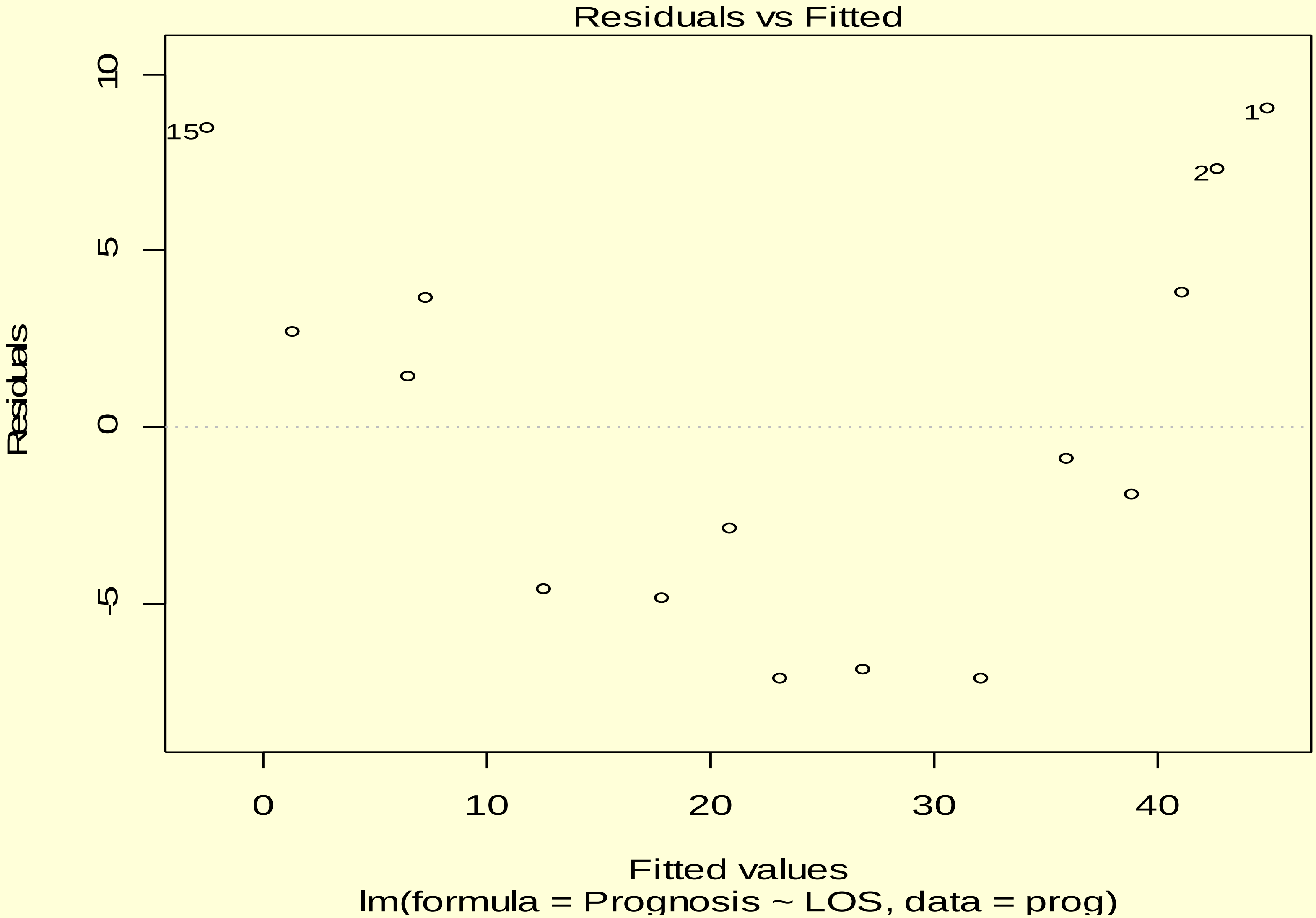

0.4
0.2
0.0
-0.2
-0.4
resid(lm(b ~ a + d))
0
1
2
3
4
5
fitted(lm(b ~ a + d))

Another Quick Example

- A hospital administrator wants to develop a prediction equation for long-term prognosis using the length of the hospital stay
- Two continuous, normally distributed variables


Prognosis
10
20
30
40
50
0
10
20
30
40
50
60
LOS

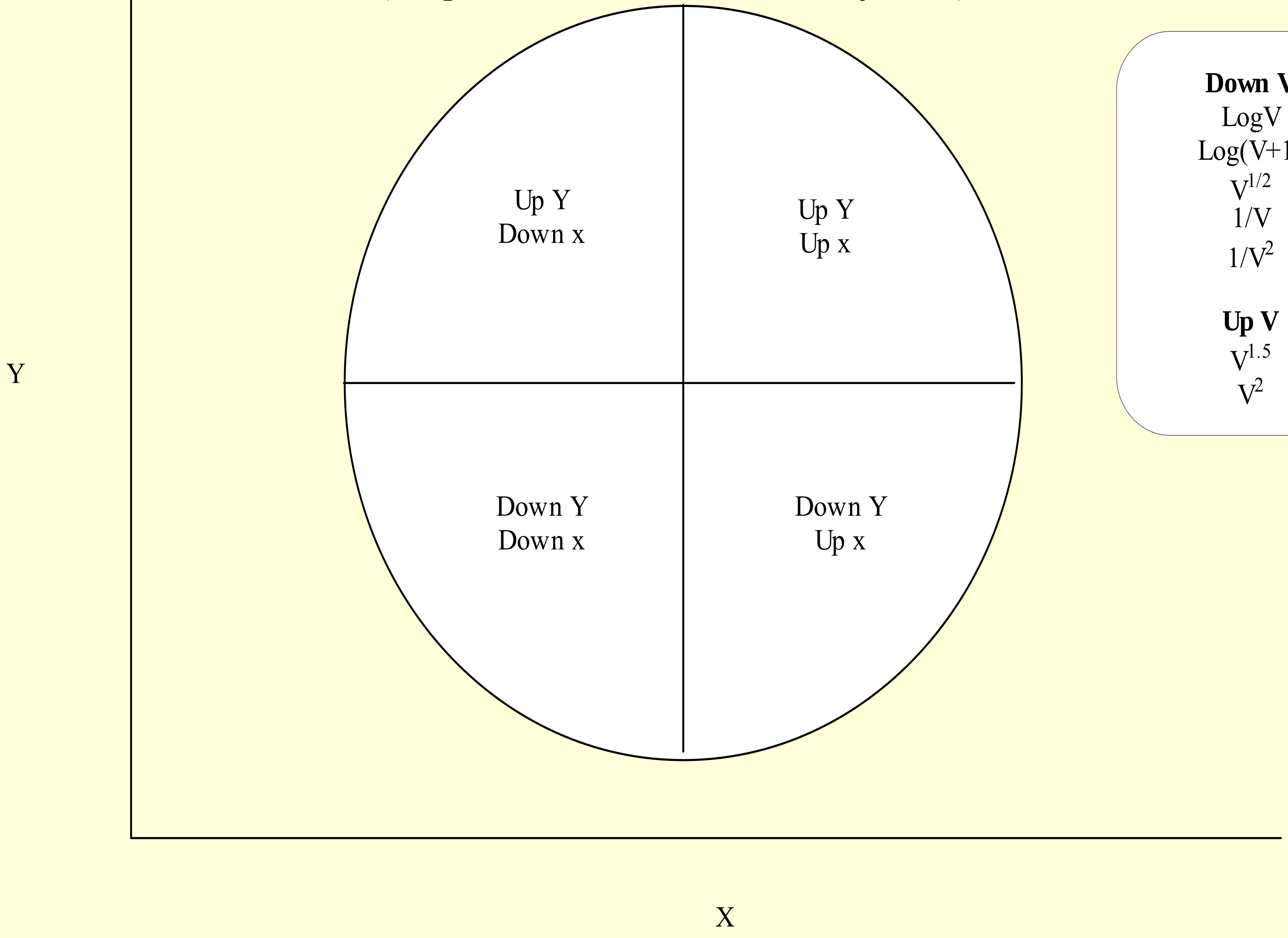
```
*** Linear Model ***


lm(formula = Prognosis ~ LOS)

Coefficients:

               Value Std. Error  t value Pr(>|t|)

(Intercept)  46.4604   2.7622    16.8202   0.0000

        LOS  -0.7525   0.0750   -10.0308   0.0000


Adjusted R-Squared: 0.8856
```


Residuals vs Fitted
Residuals
10
5
0
-5
0
10
20
30
40
Fitted values
lm(formula = Prognosis ~ LOS, data = prog)
15
2
1

How to Fix

- Stay with the linear regression framework and
  - Add a regressor which is a nonlinear function of one of the other variables
  - Apply a *nonlinear transformation* to the dependent and/or independent variables (Rule of the Bulge)
- Use a different parametric model
- Use a distribution-free strategy

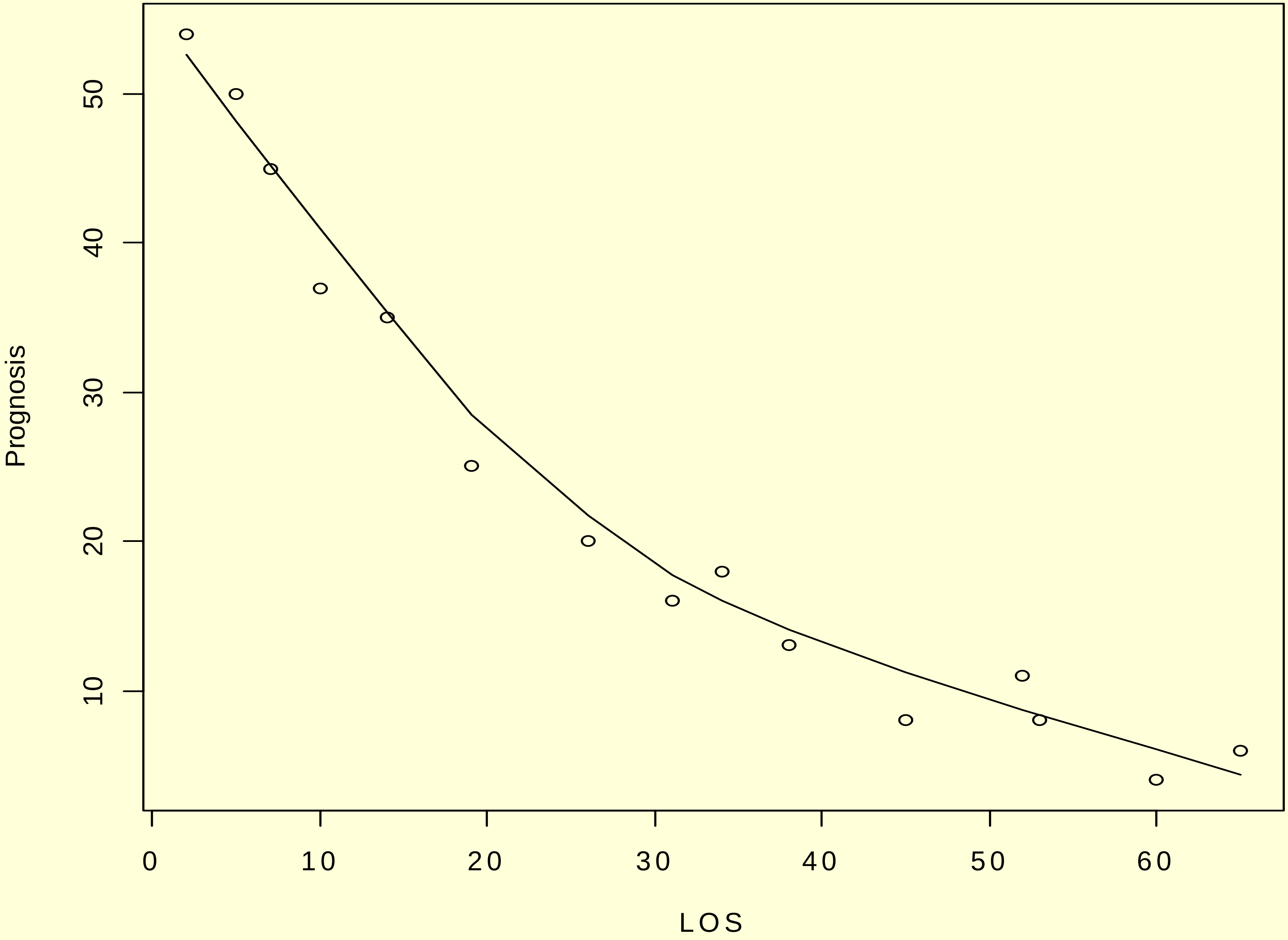
Straightening Non-linear Relationships
Rule of the Bulge
(Adapted from Mosteller and Tukey, 1977)



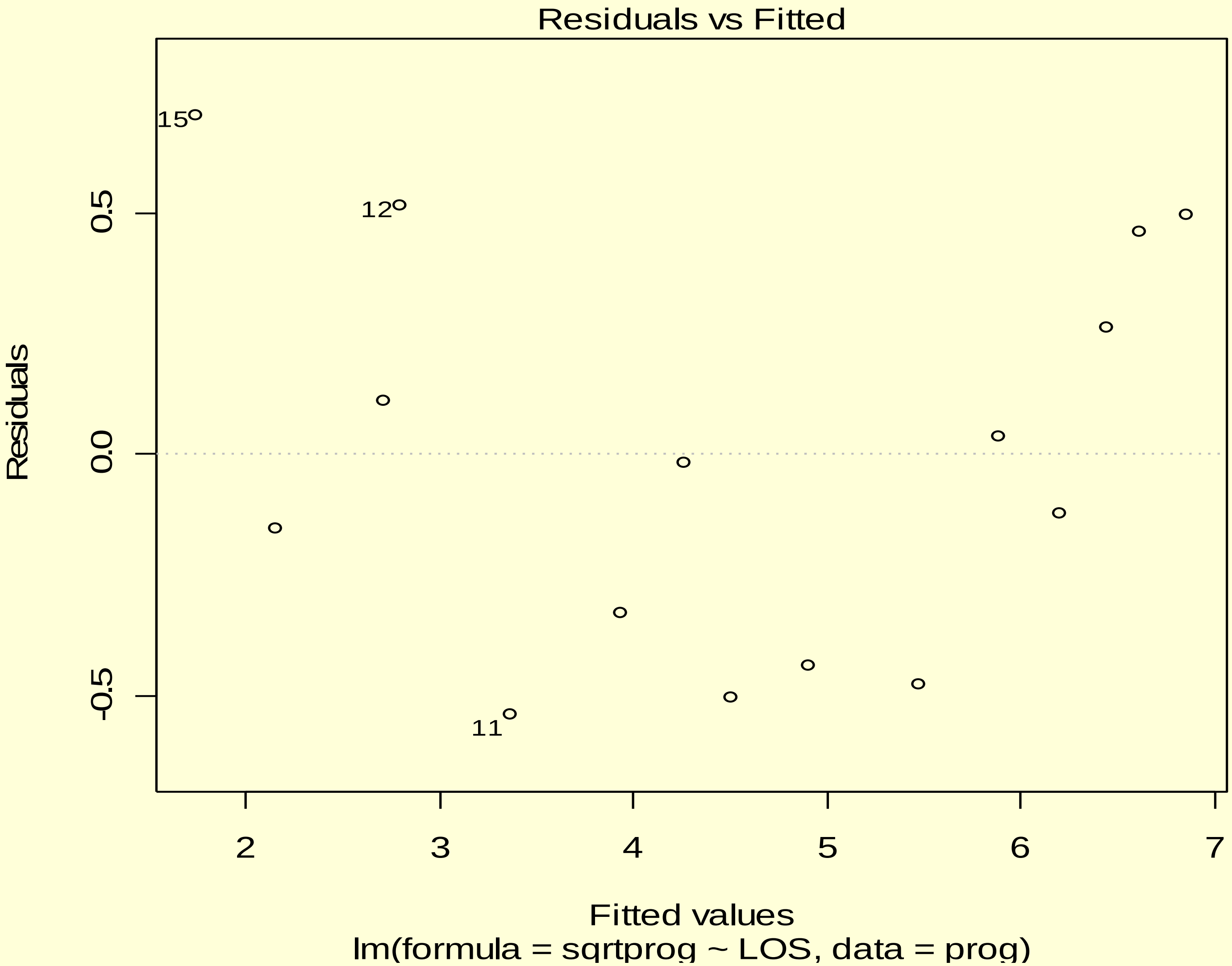
Down V
LogV
Log(V+1)
$V^{1/2}$
1/V
$1/V^2$

Up V
$V^{1.5}$
V^2

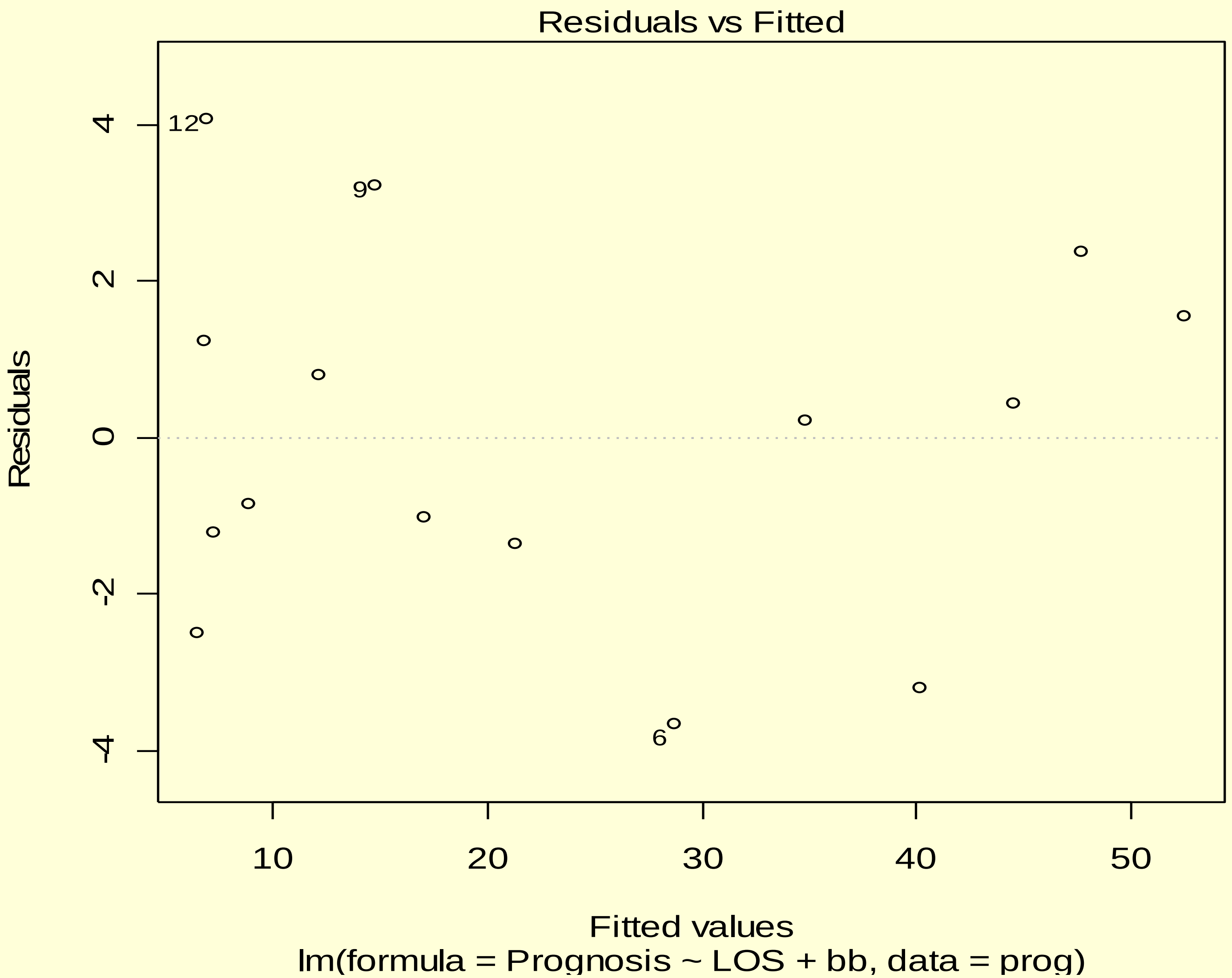
Association Between Prognosis and Length of Stay (LOS)



```
lm(formula = sqrtprog ~ LOS)

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  7.015455   0.201677   34.79 3.24e-14 ***

LOS         -0.081045   0.005477  -14.80 1.63e-09 ***

Adjusted R-squared: 0.9396
```


Residuals vs Fitted
Residuals
0.5
0.0
-0.5
15
12
11
2
3
4
5
6
7
Fitted values
lm(formula = sqrtprog ~ LOS, data = prog)

```
      *** Linear Model ***


lm(formula = Prognosis ~ LOS + LOS^2)

Coefficients:

               Value Std. Error  t value Pr(>|t|)

(Intercept)  55.8221   1.6492    33.8480   0.0000

        LOS  -1.7103   0.1248   -13.7044   0.0000

    (LOS^2)   0.0148   0.0019     7.9273   0.0000


Adjusted R-Squared: 0.9817
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
12
9
6
10
20
30
40
50
Fitted values
lm(formula = Prognosis ~ LOS + bb, data = prog)

Violations of the Normality Assumption

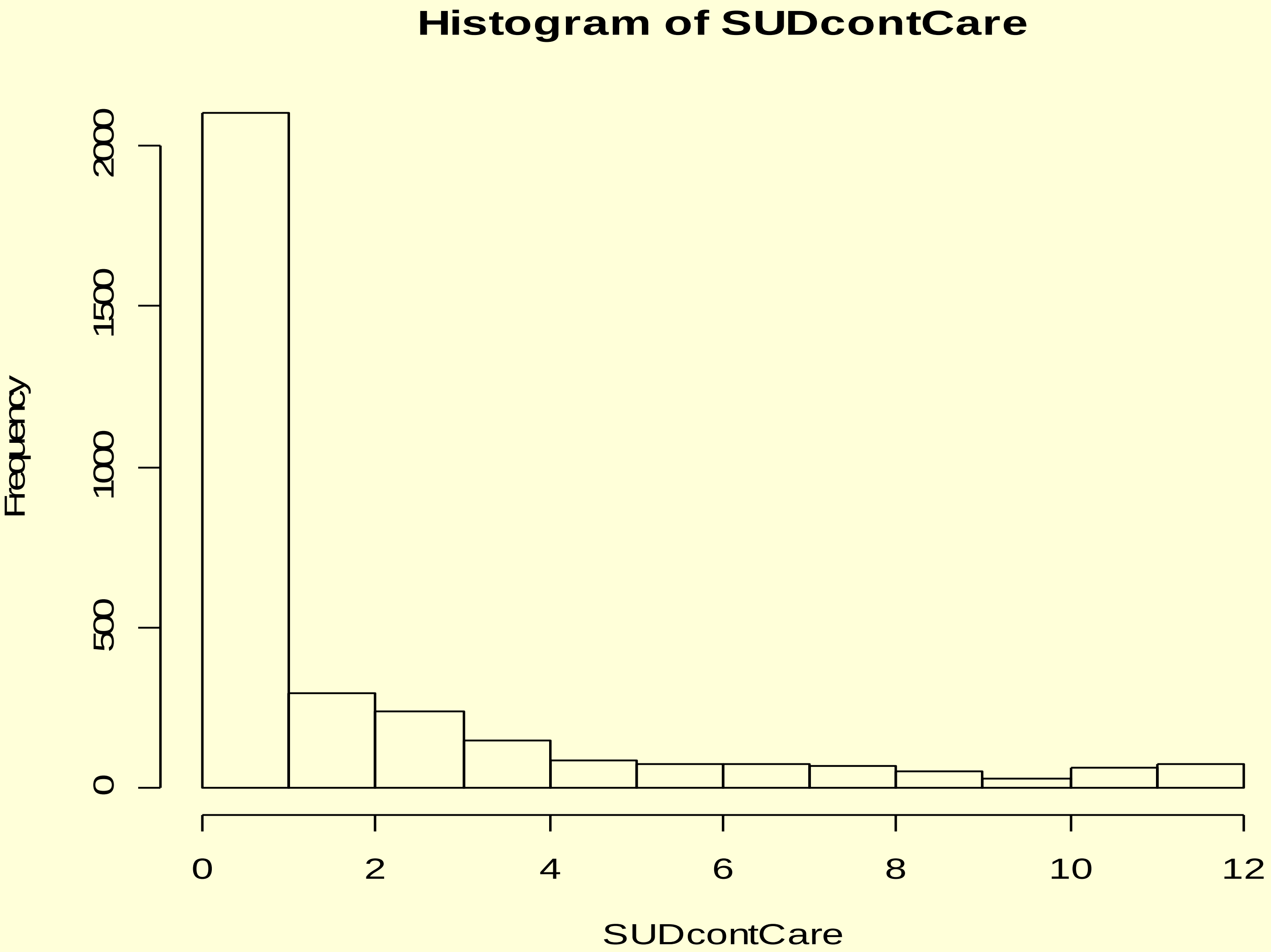
- Calculation of confidence intervals and significance tests for coefficients are based on the assumptions of normally distributed errors.
- If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

Common Causes

- The *linearity assumption* is violated
- The *distributions of the dependent and/or independent variables* are significantly non-normal

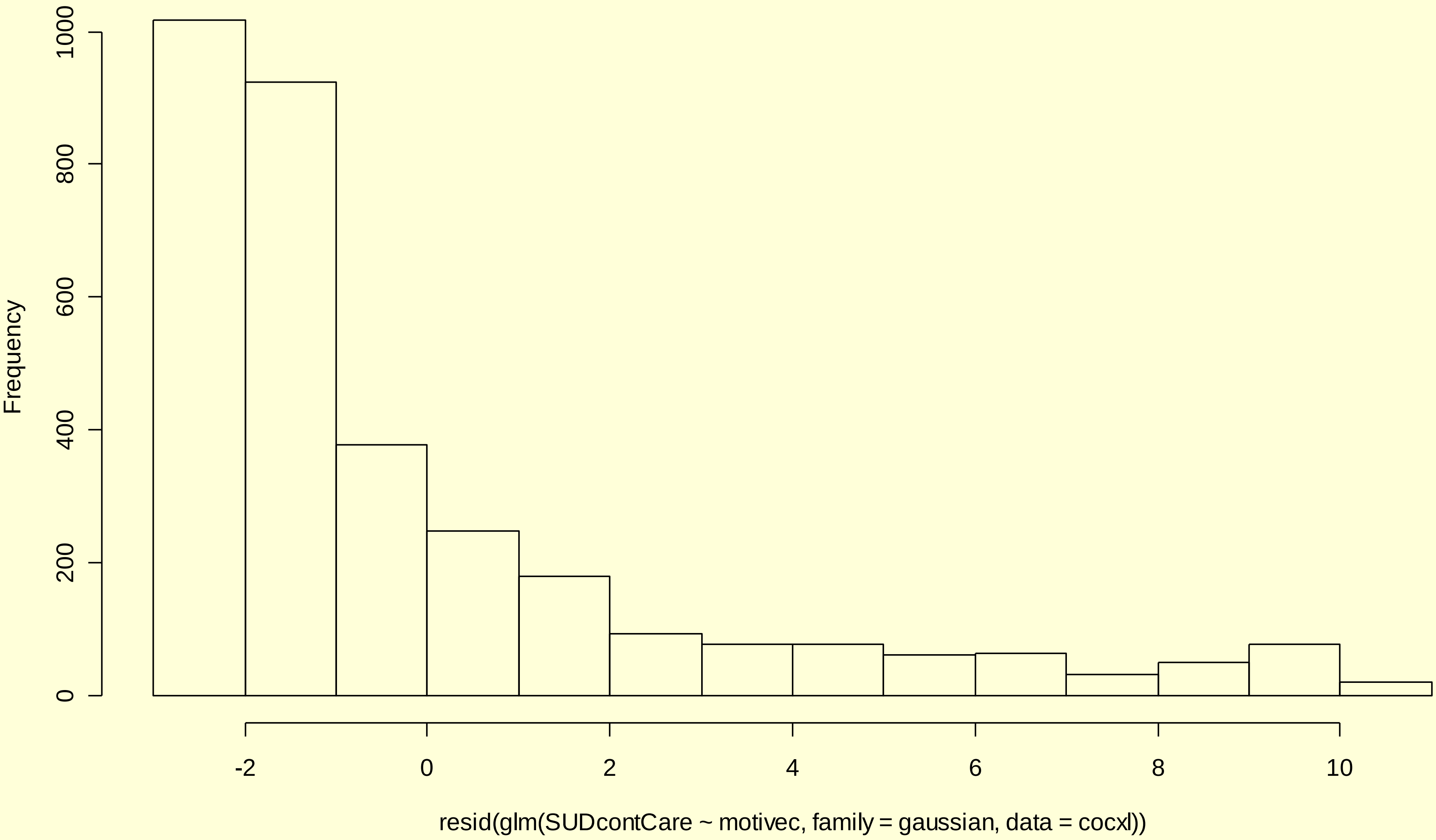
How to Detect Violations of Normally Distributed Errors

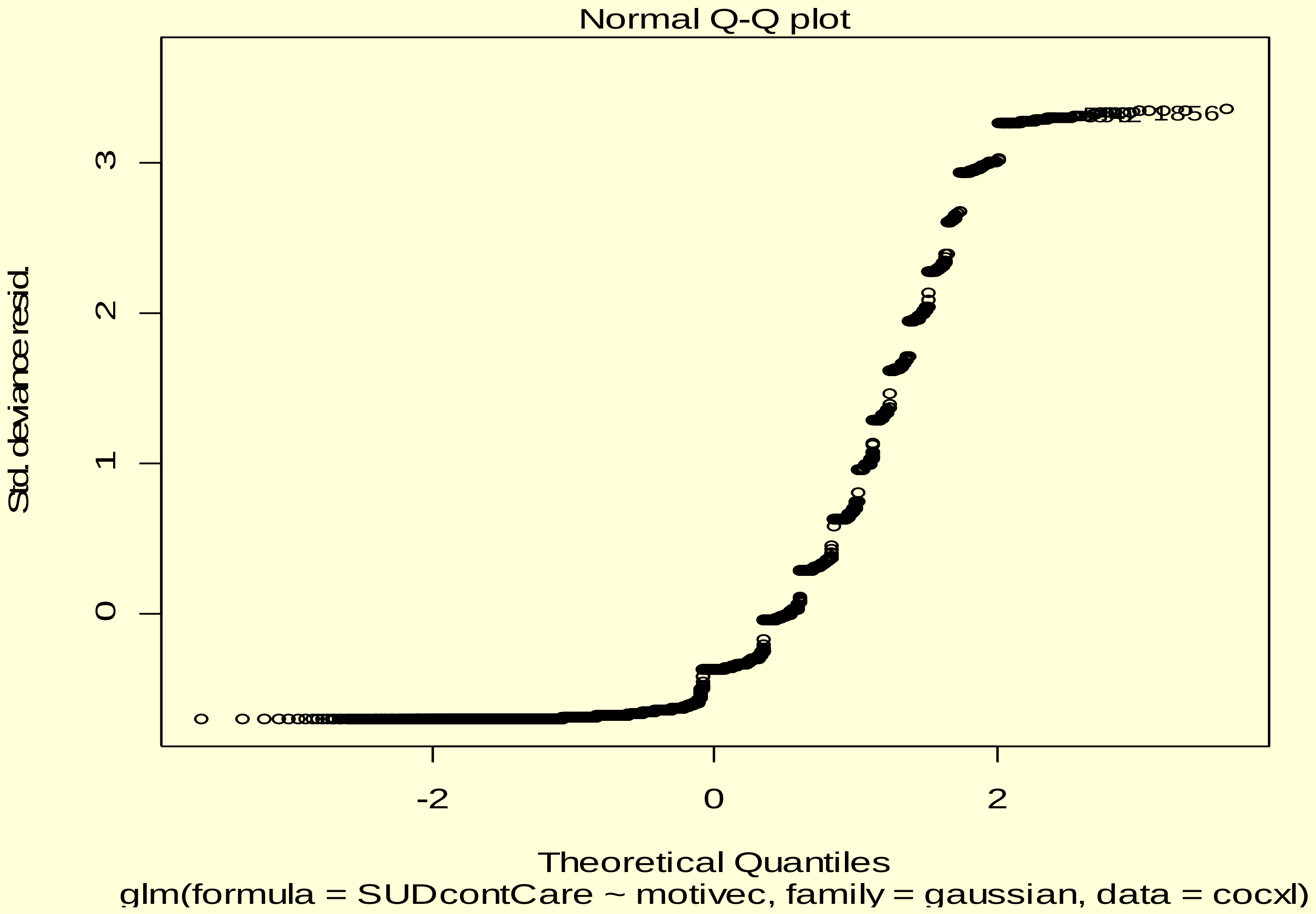
- *Histogram of the residuals*
- *Normal (QQ) probability plot* of the residuals. Quantiles of error distribution versus the quantiles of a normal distribution.


Histogram of SUDcontCare
Frequency
0
500
1000
1500
2000
0
2
4
6
8
10
12
SUDcontCare

```
glm(formula = SUDcontCare ~ motivation, family =
gaussian)

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  2.02292    0.05290  38.243   <2e-16 ***

motivation   0.02725    0.01348   2.021   0.0433 *
```

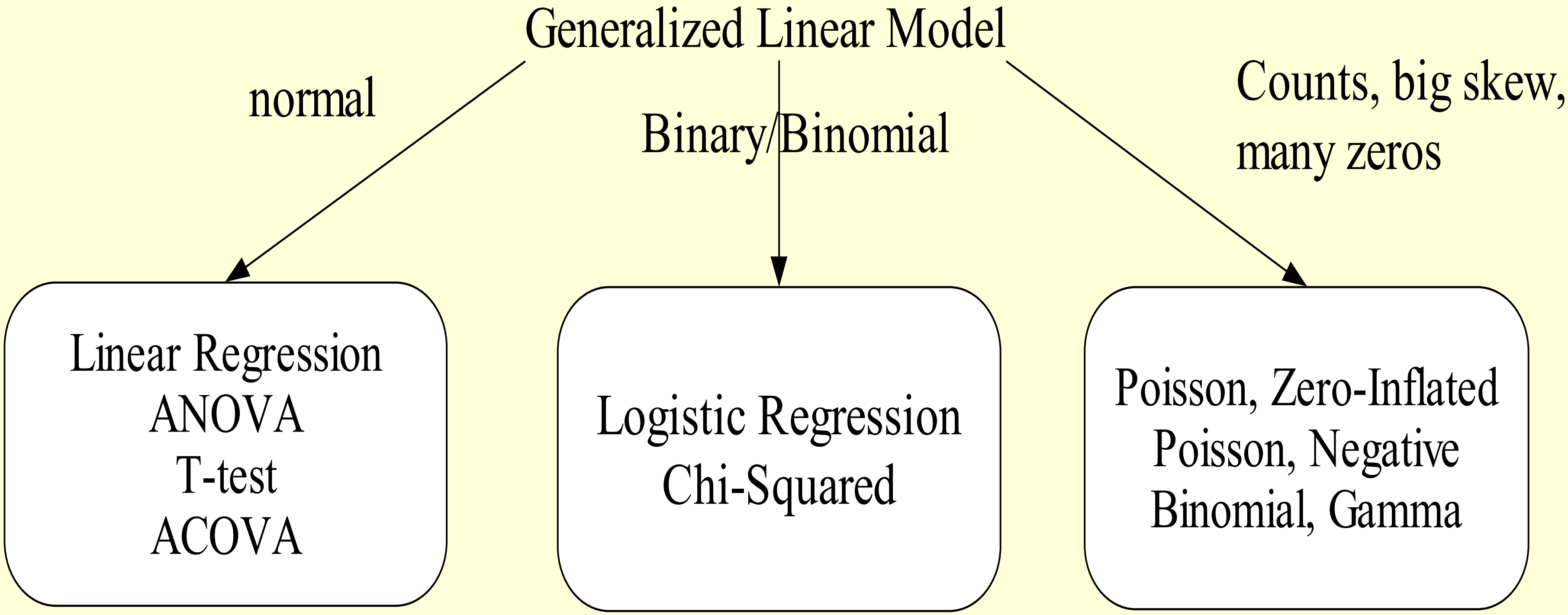
Histogram of resid(glm(SUDcontCare ~ motivec, family = gaussian, data = cocxl))




Normal Q-Q plot
Std. deviance resid.
3
2
1
0
-2
0
2
1856
Theoretical Quantiles
glm(formula = SUDcontCare ~ motivec, family = gaussian, data = cocxl)

How to fix Violations of Normally Distributed Errors

- Fix non-linearity problems
  - – Transformations
  - – Scrutinize outliers
  - – Use a non-linear model (e.g., polynomial)
- Use a different Parametric Model that fits your data
  - – Poisson Regression (counts of rare events)
  - – Negative Binomial (often better with overdispersion)
- Use a non-parametric Model
  - – bootstrap



Generalized Linear Model

normal

Binary/Binomial

Counts, big skew,
many zeros

Linear Regression
ANOVA
T-test
ACOVA

Logistic Regression
Chi-Squared

Poisson, Zero-Inflated
Poisson, Negative
Binomial, Gamma

```
glm(formula = SUDcontCare ~ motivation, family =
poisson)

Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  0.70303    0.01227  57.281  < 2e-16 ***

motivation   0.01410    0.00327   4.314 1.60e-05 ***
```

Note: In terms of proportion of deviance “explained”, this model is 25% better than the Gaussian model

Non-parametric, Distribution-free (not assumption free) Methods

- Ranks, signs, etc
- Bootstrap Regression
  - Estimate the distribution of the parameters by resampling with replacement.
  - Construct confidence intervals based on the empirical distribution

Summary Part 1

- Know the assumptions of the models you use
- Check for Model/Data incongruence
- If incongruence is found, adjust your data and/or your model
- Know your options

Suggested Resources for Learning About Linear Models and Alternatives

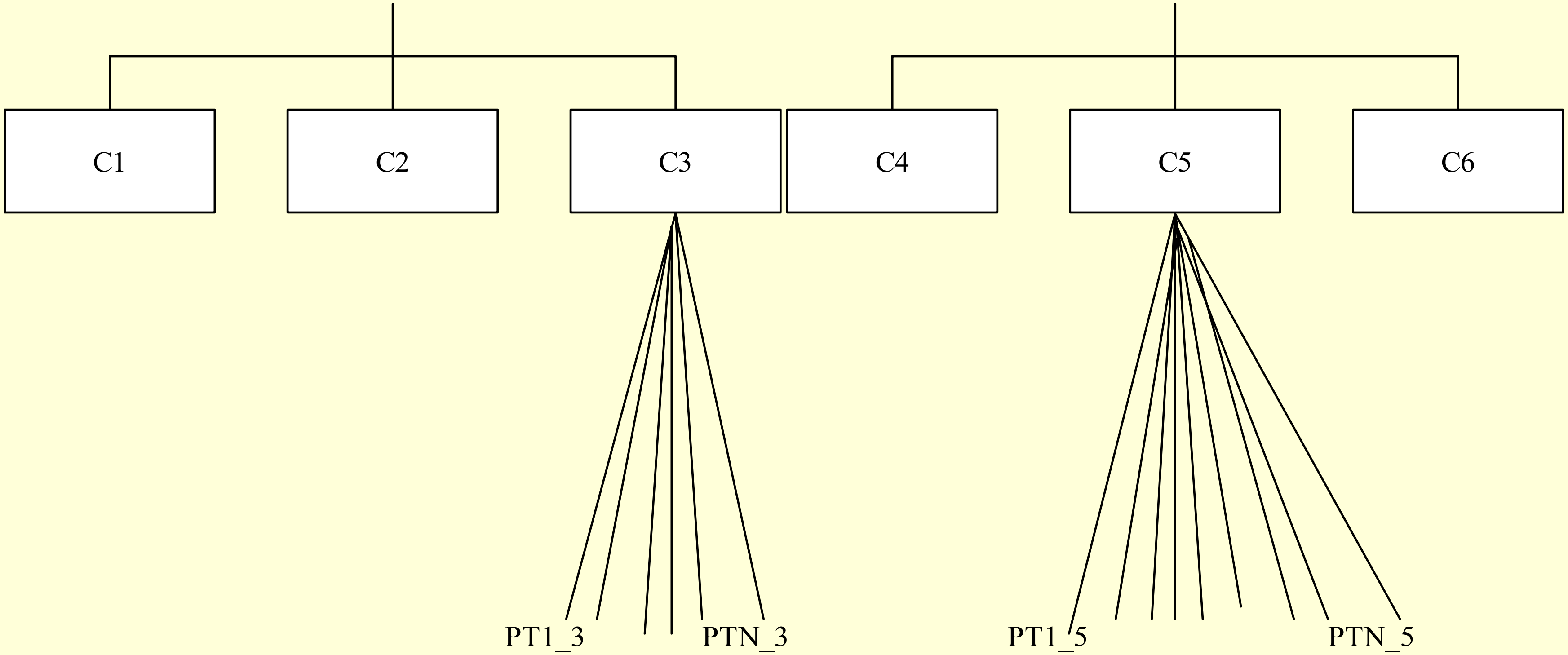
- Ramsey, F. L., & Schafer, D. W. (1997). *The Statistical Sleuth*. Belmont, CA: Duxbury Press.
- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Montgomery, D., & Peck, E. (1992). *Introduction to Linear Regression Analysis* (2nd ed.). New York, NY: Wiley.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models* (4th ed.). Chicago: Irwin.
- Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall.

Part 2: Correlated Data:
When Errors are Not Independent

Common Data Structures

- Multi-Level Organizational Data
  - Patients within providers within facilities

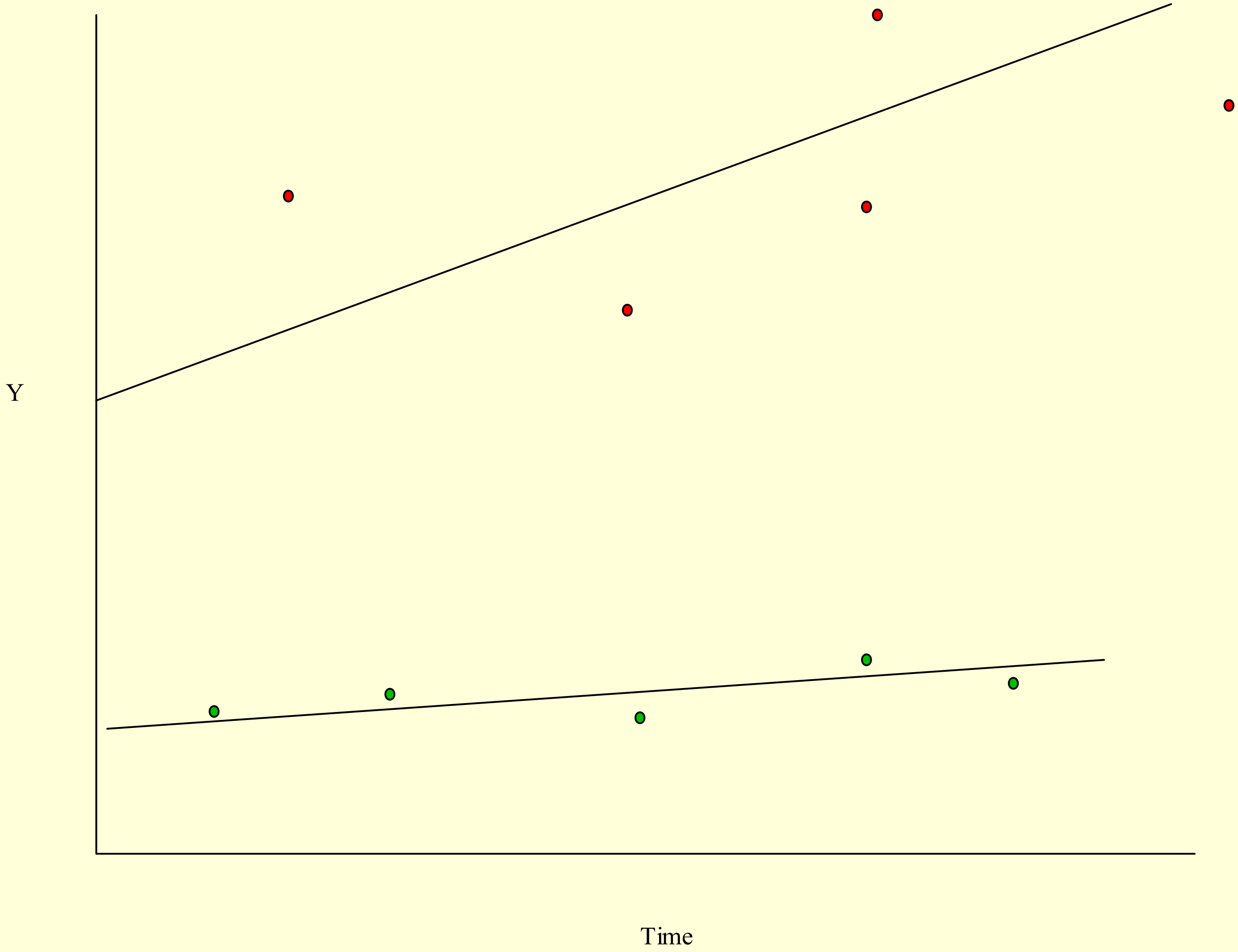
Multi-Level Organizational Data



Common Data Structures

- Repeated-Measures on Individuals
  - Monthly measurement of disease status

Repeated-Measures on Individuals



Common Data Structures

- Both within person and within organization clustering

The Problem

- Common statistical tools have no good way of dealing with multi-level details (correlated errors, sample size, variances)
  - OLS Regression
  - ANOVA
  - T-tests
- It matters – failing to attend to these details can give very wrong results.

Old (and usually bad) Solution

- To aggregate or disaggregate data to one level and apply familiar statistical models.

Example

- Study: What are the clinic characteristics (e.g., open on weekends) that influence patient outcomes?
  - Sample is 700 patients in 20 clinics

- Force all information to the patient-level
  - Confounds patient and clinic sample sizes
  - Radically reduced the SE of parameter estimates
  - Leads to more null-hypothesis rejection and inappropriately narrow CIs

- Force all information to the Clinic-level
  - Lose power
  - Lose information about within clinic variability and size

Mixed Effects Regression Keeps Track of Multi-level Details and Allow for Dependencies

- Allows you to
  - address single-level questions while accounting for dependencies at other levels.
  - appropriately test interesting and important multi-level hypotheses.
    - Clinic characteristics that influence patient outcomes

Punch Line 2

- Multi-level thinking is powerful and important conceptually and statistically
- Need to use models that keep track of multi-level details
  - Sample size
  - Variance partition
  - Correlated errors

Example: Cross-Level Question

- 2931 SUD patients in 15 clinics
- Clinics have been rated for level of guideline concordant care (0-10)
- Question: Does guideline concordant care influence patient engagement in continuing care (“aftercare”; also a normally distributed variable)

Analysis that Confounds Clinic and Patient Sample Size

```
glm(formula = Engage ~ Concord, family = gaussian)

Coefficients:
            Estimate Std. Error t value Pr(>|z|)
(Intercept)  -1.4062     0.1567  -8.973   <2e-16 ***
Concord       0.2311     0.0221  10.457   <2e-16 ***
```

Variance Partitioning in Regular Regression

$$y_i = \beta_0 + \beta_1 (CONCORD_i) + e_i$$

$$\text{where } e_i \sim N(0, \sigma^2)$$

Variance Partitioning in Multi-Level Regression

$$y_{ij} = \gamma_{00} + \gamma_{01}(\mathrm{CONCORD})_j + \mu_j + e_{ij}$$

$$\text{where } \mu_j \sim N(0, \sigma^2_{\mu}), \quad e_i \sim N(0, \sigma^2_e)$$

Analysis that Keeps Track of Levels

```
lme(ENGAGE ~ CONCORD , random  = ~ 1  | SITE, family =
gaussian)

Random effects:
 Formula: ~1 | SITE
        (Intercept) Residual
StdDev:   0.7021689 2.036858

Fixed effects: ENGAGE ~ CONCORD
                 Value Std.Error   DF    t-value p-value
(Intercept) -1.4627475  1.609469 2916 -0.9088384  0.3635
CONCORD    0.2145252  0.225683   13  0.9505596  0.3592
-----------------------------------------------------------
Concord      0.2311     0.0221           10.457   <2e-16 ***
```

Example: Level-1 Question with Level-2 Grouping

- 2931 SUD inpatients in 15 clinics
- Question: Do SUD inpatients’ ratings of staff control influence engagement in continuing care?
- Note: Patients ratings within each site are likely to be correlated.

Not Accounting for Clustering within Clinics

```
Call:
glm(formula = ENGAGE ~ PRCONTROL, family = gaussian)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.13774    0.08450  -1.630    0.103
PRCONTROL    0.04965    0.01178   4.214 2.50e-05 ***
```

Accounting for Clustering within Clinics

lme(ENGAGE ~ PRCONTROL , random = ~ 1 | SITE, family = gaussian)

```
Random effects:
 Formula: ~1 | ISITE
        (Intercept) Residual
StdDev:   0.7253934 2.037000
```

Rho = ICC = 0.11 → variance explained by the grouping

```
Fixed effects: Engage ~ PRCONTROL
                  Value  Std.Error   DF    t-value p-value
(Intercept)  0.11391253 0.27017976 2915  0.4216176  0.6733
PRCONTROL   -0.00823548 0.02688099 2915 -0.3063682  0.7593

Number of Observations: 2931
Number of Groups: 15
---------------------------------------------------------
PRCONTROL    0.04965    0.01178           4.214  2.50e-05 ***
```

Mixed Effects Regression Models, aka HLM, random coefficients models, growth curve models, etc.

- A very flexible framework to handle multi-level data and questions.
- Allow various link functions (normal, binomial, possion, etc)
- Can explicitly model the covariance structure
- Handle unbalanced data and variable assessment schedules
- All cases can be included

Multi-level Models for Longitudinal Data

- Observations are clustered within individuals
- Individual outcome trajectories are modeled over time (intercept and slope)
- Questions about intercept/slope relationship and the effect of predictors on these.

RCT Example

- Effect of patent-level intervention on trajectories of perceived stress

- 252 patients randomized to active treatment or usual-care control

- Perceived stress measured at baseline, post-intervention (6-weeks) and 20 weeks.

Modeling Individual Trajectories Over Time (within-person = Level 1)

$$Y_{it} = \pi_{0i} + \pi_{1i}(time)_{it} + e_{it}$$

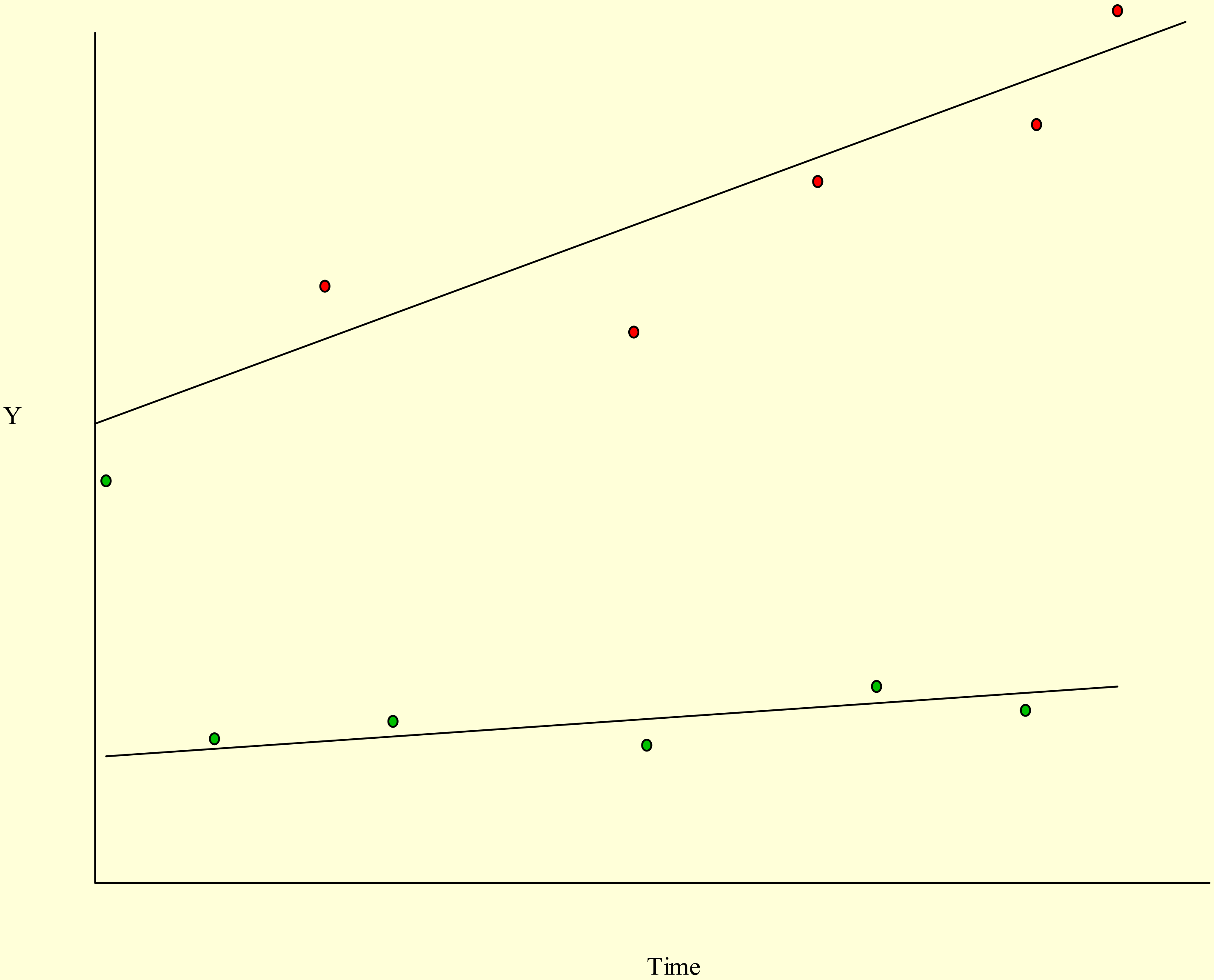
Y_{ti} is the outcome at time t for patient i

π_{0i} is the initial status of patient i

π_{1i} is the rate of change for patient i

$(time)_{it}$ is 0 at intake

e_{it} is the error associated with patient i at time t


Y
Time

Level 2: Between-Person Effects

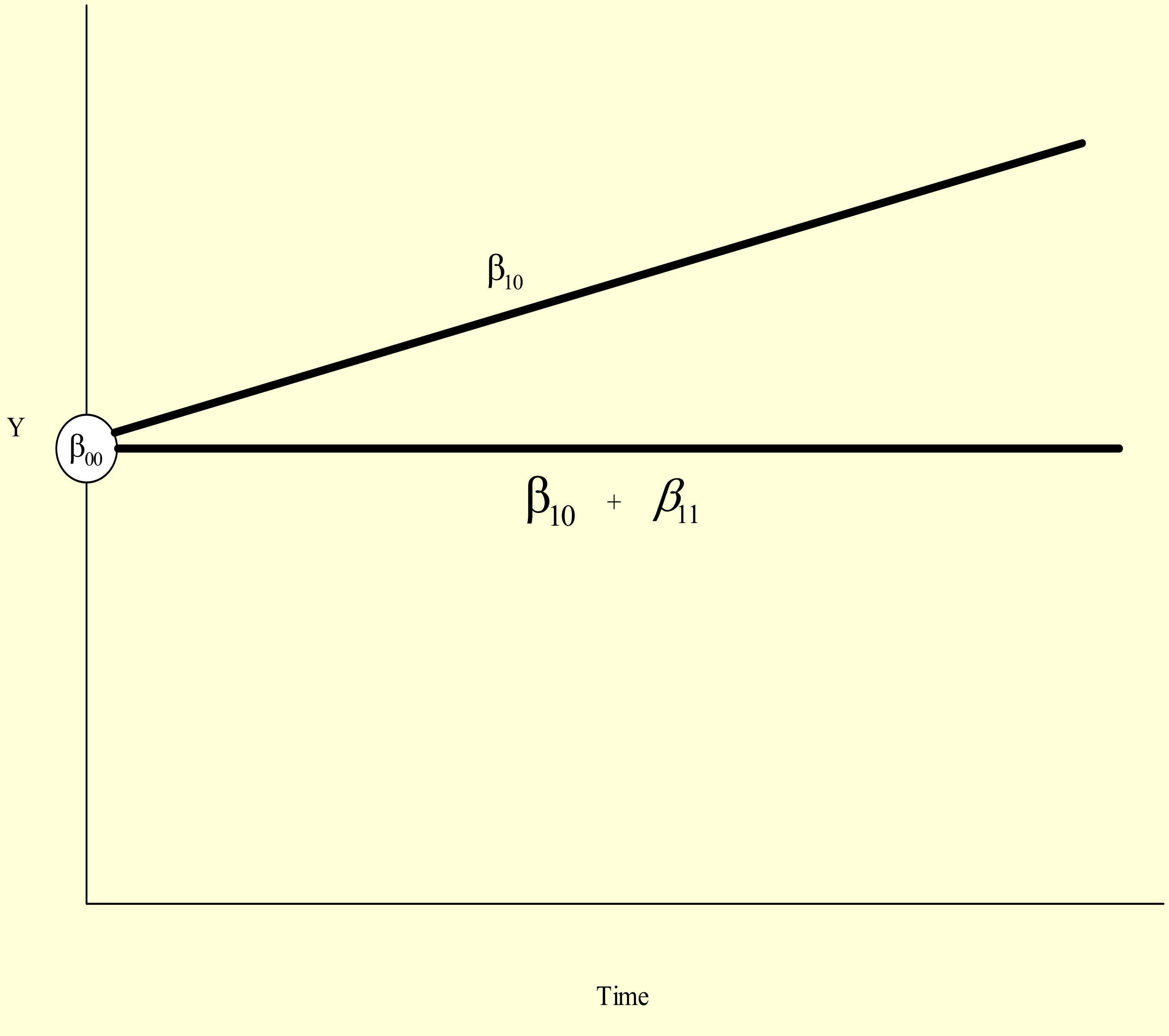
$$\pi_{0i} = \beta_{00} + r_{0i}$$

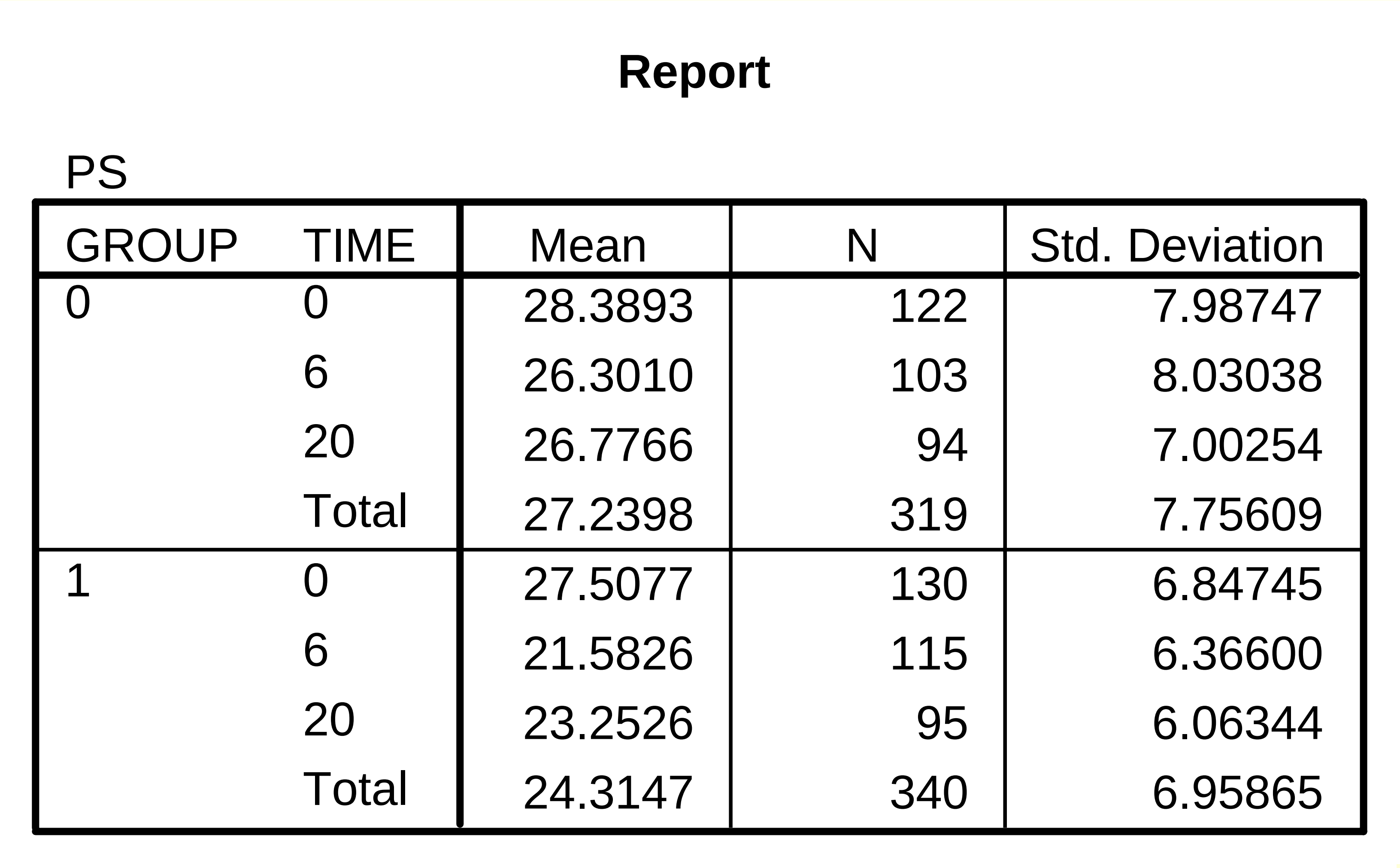
$$\pi_{1i} = \beta_{10} + \beta_{11}(\mathrm{TreatmentGroup})_i + r_{1i}$$

β_{00} is the average initial status for all patients

β_{10} is the average slope for control group

r_{0i} and r_{1i} are errors associated with person i


Y
β_{00}
β_{10}
$\beta_{10} + \beta_{11}$
Time

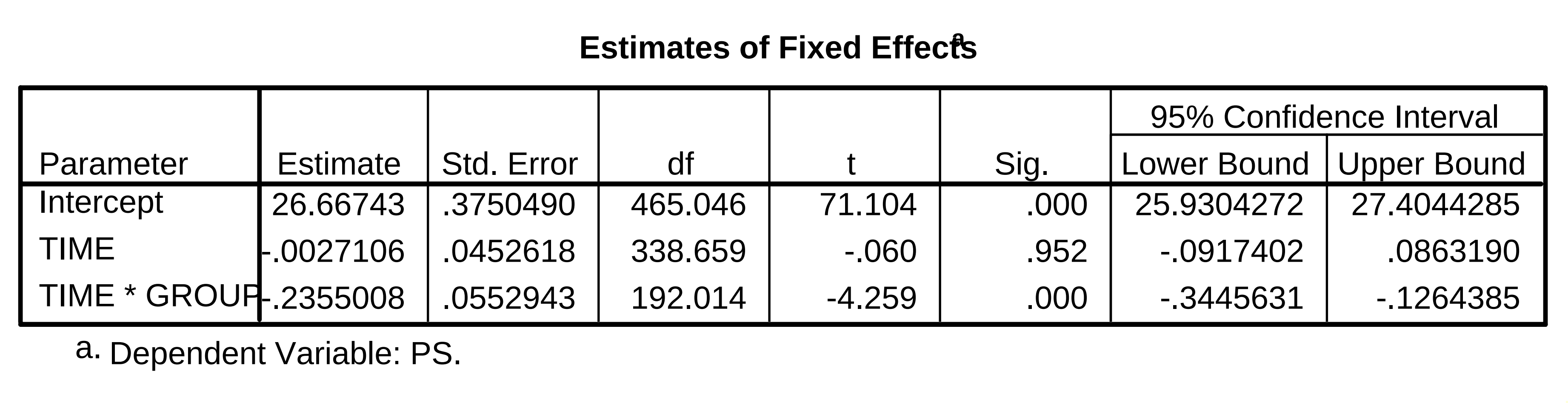


Report

PS

GROUP	TIME	Mean	N	Std. Deviation
0	0	28.3893	122	7.98747
	6	26.3010	103	8.03038
	20	26.7766	94	7.00254
	Total	27.2398	319	7.75609
1	0	27.5077	130	6.84745
	6	21.5826	115	6.36600
	20	23.2526	95	6.06344
	Total	24.3147	340	6.95865

```
mixed ps with time group
  /print=solution
  /method=ml
  /fixed = intercept time time*group
  /random intercept time | subject(id) covtype(ar1).
```



Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	26.66743	.3750490	465.046	71.104	.000	25.9304272	27.4044285
TIME	-.0027106	.0452618	338.659	-.060	.952	-.0917402	.0863190
TIME * GROUP	-.2355008	.0552943	192.014	-4.259	.000	-.3445631	-.1264385

a. Dependent Variable: PS.

Repeated Measures ANOVA

- Requires complete data
- Does not keep track of individuals
- Time*Treatment interaction F-test is often a poor operationalization of our research questions

Suggested Resources for Learning About Mixed Effects Models

- Pinheiro, J., & Bates, D. (2000). *Mixed Effects Models in S and S-Plus*. New York, NY: Springer.
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models : Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. London: Oxford University Press.
- Hox, J. J. () Applied Multilevel Analysis. http://www.fss.uu.nl/ms/jh/publist/amaboek.pdf

Suggested Resources for Learning About Mixed Effects Models

- Another list of books http://www.ats.ucla.edu/stat/books/#Multilevel
- UCLA’s Multi-level Modeling Portal http://statcomp.ats.ucla.edu/mlm/default.htm
- Software Reviews of Multilevel Analysis Packages http://www.mlwin.com/softrev/index.html

Covariance Structure Specification

- Can not only stipulate that data are correlated within specific units, but you can model this correlation
  - Unstructured
  - Autoregressive
  - Exchangeable

Mixed Effects Regression

- Fixed and Random variables – issue of error assumptions
- Fixed and Random Effects – issues of generalizability
- Fixed Coefficients and Random Coefficients
- We want to estimate the effect of predictor on outcomes that generalize beyond the particular groups in the study
- http://www.upa.pdx.edu/IOA/newsom/mlrclass/ho_randfixd.doc

Missing Data

Part 3

Goals

- Advantages and disadvantage of various strategies for addressing missing data.
- Discuss criteria for evaluating methods and specifying “best-practices.”

The Punch Line

Choices regarding how you handle missing data may dramatically affect the accuracy, efficiency (power), and reliability of the inferences you make.

The Purpose of Analysis

- To make valid inferences regarding a population of interest.
- To estimate the population means, variances, inter-correlations, and error for these estimates = multivariate parameter space θ
- Not to guess what a particular missing value would have been.

Criteria for Choosing a Method

- Allow us to make valid inferences to the population of interest
- Account for different status of observed and missing values.
- Convenient :
  - Easy to implement
  - Requires no special software

Properties of Good Missing Data Methods (Heitjan & Little, 1991)

- Should condition on observed variables for that case.
- Should account for the multivariate nature of non-response (consider the overall distribution of missingness)
- Should not distort the marginal distributions and associations in the complete (observed and missing) data

$$Y_{complete} = Y_{obs} + Y_{mis}$$

Evidence of Goodness and Badness of Methods: Simulation Studies

- Start with a population and select a sample
- “erode” the data into various patterns of missingness.
- Apply an approach to addressing the missingness.
- Compare estimates to known population values

Common Approaches to Missing Data

- Throw away cases with missing values- act as if they had never been observed

- Fill-in missing data using some method (e.g., mean or regression imputation), and then treat them as if they had always been observed.

Taxonomy of Missing Data Methods

- Omit Cases
- Imputation
  - Single and multiple
  - Good and bad methods
- Model-Based Approaches
  - Define a model for the complete (missing+observed) data and use maximum likelihood or iterative simulation to estimate population parameters

Ways Data Can be Missing Rubin (1976)

- Missing Completely at Random (MCAR)
  - Missing values are a random sample of all values. Missingness does not depend on Xs or Y
- Missing at Random (MAR)
  - Missing values can depend on the value of observed variables (Xs) but not on the values of Y.
- Missing Not at Random (MNAR)
  - Missingness depends on value of Y even after relationships with Xs have been accounted for.

Tour de Missing Data Methods

- Complete Case Analysis

Complete Case Analysis (aka listwise deletion)

- Do analysis using complete cases. Treat omitted data like they never existed.
- Strengths
  - Easy
  - Certain intuitive appeal
  - Is unbiased and efficient in very restricted circumstances (univariate MCAR)

Complete Case Analysis (Aka listwise deletion)

- Limitations
  - To the extent that completely observed cases differ from cases with missing data (MAR), can create serious bias.
  - Inferences are then valid for completers rather than the intended population.
  - Fails to keep track of uncertainty due to missingness.

Example: SBP Data

- Two variables:
  - X = SBP in Jan
  - Y = SBP in Feb
- 1000 datasets of size 50 (complete data) drawn from a known parameter distribution.
- Erode data (Y) into MAR, MCAR, MNAR.
- 75% Y missing
- See how well methods perform under these conditions

Performance of LD for Parameter Estimates (Schafer &Graham, 2002)

Parameter	MCAR	MAR	MNAR
$\mu_Y = 125.0$	125.0	143.3	155.5
$\sigma_Y = 25.0$	24.6	20.9	12.2
$\rho = .60$	.59	.33	.34

Tour de Missing Data Methods

- Complete Case Analysis
- Available Case Analysis

X_j

Available Case Analysis (aka pairwise deletion)

- Use all available data to estimate parameters of interest. May use different number of cases to estimate various parameters.
- Use every observed value of x to calculate SD(x), every observed pair (x,y) to calculate cov(x,y).

Available Case Analysis (aka pairwise deletion)

- Strengths same as LD but better because it uses more of the data.
- Weaknesses similar to LD
  - Most notably, may be seriously biased
  - Because parameters are estimated from different cases, difficult to estimate SEs and other measures of uncertainty.

Tour de Missing Data Methods

- Complete Case Analysis
- Available Case Analysis
- Mean Imputation

Mean Imputation

- Replace missing values with the mean for that variable
- Strengths
  - Easy, built into SPSS
  - Certain intuitive appeal
  - Does not bias means

Mean Imputation

- Limitations
  - Creates bias in variances and covariances (toward zero). Does not condition on observed data for each case
  - Fails to keep track of uncertainty due to missingness.

Performance of Mean Substitution for Parameter Estimates (Schafer &Graham, 2002)

Parameter	MCAR	MAR	MNAR
$\mu_Y = 125.0$	125.1	143.5	155.5
$\sigma_Y = 25.0$	12.3	10.6	6.20
$\rho = .60$	.30	.08	.15

Tour de Missing Data Methods

- Complete Case Analysis
- Available Case Analysis
- Mean Imputation
- Regression-based Imputation

Regression-based Imputation (aka Conditional Mean Imputation)

- Replace missing data with a linear combination of other values for that case.
- Strengths
  - – Easy, built into SPSS
  - – Conditions on other values for each case
  - – Certain intuitive appeal

Regression-based Imputation

- Limitations
  - May create bias in variance and covariance (away from zero).
  - Fails to keep track of uncertainty due to missingness.
  - Even if unbiased, confidence intervals fail to cover parameters

Performance of Regression-Based Imputation for Parameter Estimates (Schafer &Graham, 2002)

Parameter	MCAR	MAR	MNAR
$\mu_Y = 125.0$	125.2	124.9	151.6
$\sigma_Y = 25.0$	18.2	20.4	8.42
$\rho = .60$	.79	.64	.55

Regression-based Imputation

- Unbiased except for MNAR.
- Unbiased is good but not enough…
  - Proportion of CIs that cover the population parameter is an important criteria
- Coverage for LD, CD, MS, and CMS are really bad (mostly less than 50%)
- This is why single imputation strategies are dangerous.

Tour de Missing Data Methods

- Complete Case Analysis
- Available Case Analysis
- Mean Imputation
- Regression-based Imputation
- Hot Deck Imputation

Hot Deck Imputation

- Class of methods (single or multiple imputations)
- Replace missing values with a random draw from observed values.
- A refinement replaces values with a draw from neighbors on observed values.

Hot Deck Imputation

- Strengths
  - Good at preserving means and variances
- Limitations
  - Without significant refinements, measures of association are distorted

Tour de (generally not recommended) Missing Data Methods

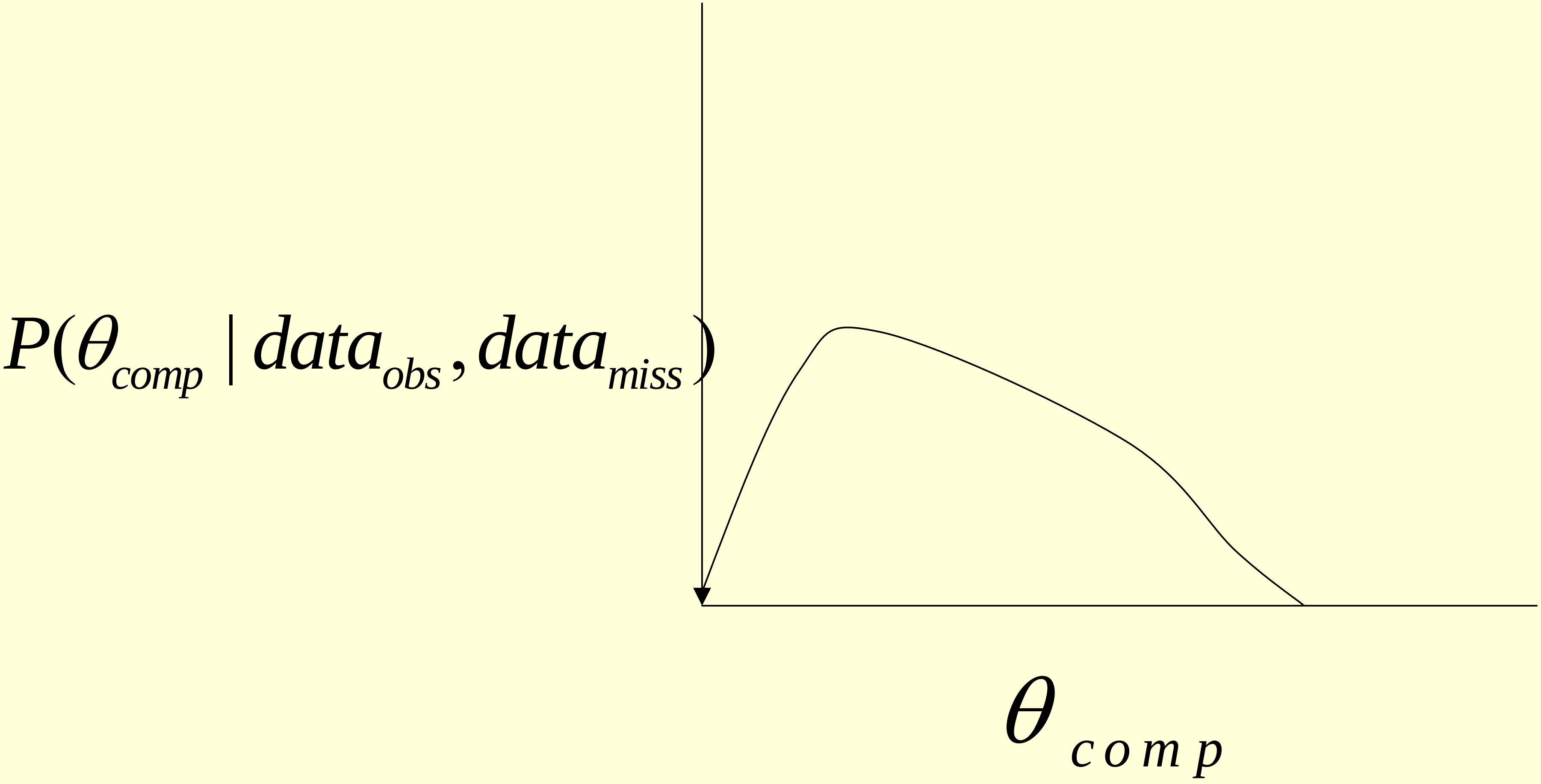
- Complete Case Analysis
- Available Case Analysis
- Mean Imputation
- Regression-based Imputation
- Hot Deck

Tour de Missing Data Methods that Have these good characteristics:

- Keep tract of uncertainty due to missingness.
- Condition on observed values for a case.
- Results in (more) accurate, efficient, and reliable inferences under MAR.
- CIs are narrow and cover the population parameter about 95% of the time.

Expectation Maximization (EM)

- Algorithm for finding maximum likelihood estimates for parametric models when data are not fully observed. Uses an iterative method to estimate the maximum likelihood of parameters of the complete data given the observed data and the pattern of missingness


$P(\theta_{comp} \mid data_{obs}, data_{miss})$
θ_{comp}

Expectation Maximization (EM)

- Strengths
  - Gives efficient and unbiased estimates of multivariate model parameters.
- Limitations
  - Not flexible
  - Requires that you work directly with model parameters

Performance of EM for Parameter Estimates (Schafer & Graham, 2002)

Parameter	MCAR	MAR	MNAR
$\mu_Y = 125.0$	124.8	125.2	151.6
$\sigma_Y = 25.0$	24.2	25.5	12.3
$\rho = .60$	.61	.52	.39

Multiple Imputation Methods

- Can use EM (or DA or FIML) as a basis to impute M separate complete datasets.
- Analysis is performed on all M datasets.
- Parameter estimates and SE are combined into one inference.

Multiple Imputation Methods

- Strengths
  - One set of imputations can be used for many analyses
  - Standard analytic tools and software can be used once you impute the data sets
  - Final inferences incorporate uncertainty due to missing data. Good CIs.
  - Highly efficient and unbiased even with small M
  - MI has better performance than EM in small samples.

Multiple Imputation Methods

- Limitations
  - – Need special software
  - – Takes time
  - – Relatively unknown

Performance of MI for Parameter Estimates (Schafer &Graham, 2002)

Parameter	MCAR	MAR	MNAR
$\mu_Y = 125.0$	124.9	125.3	151.6
$\sigma_Y = 25.0$	25.9	28.7	13.6
$\rho = .60$	.57	.45	.35

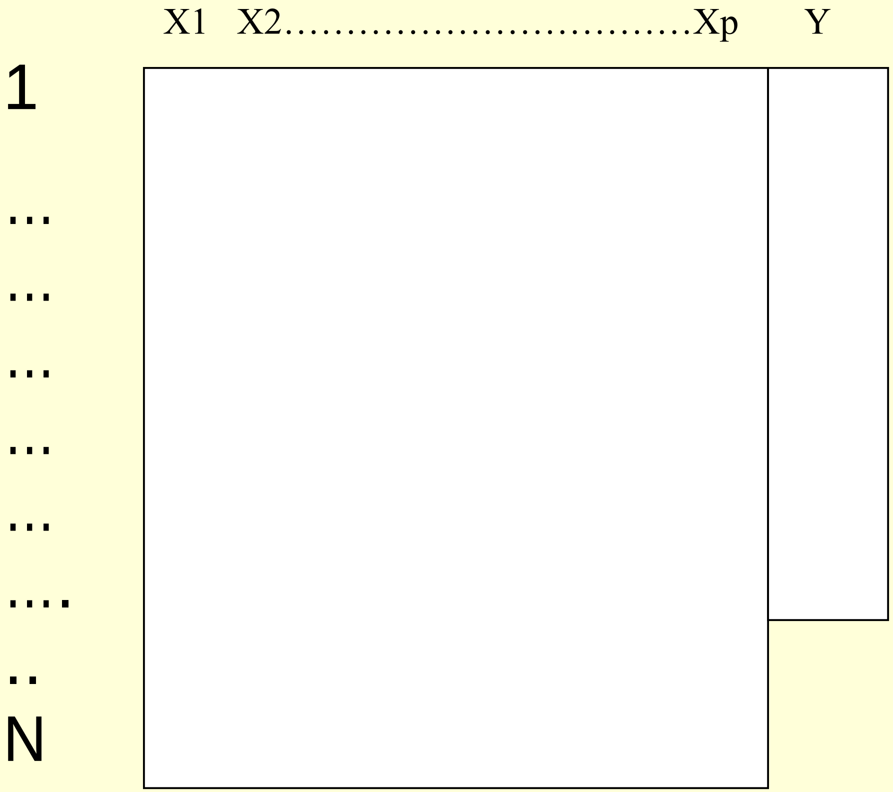
Does Proportion of Missing Data Influence Choice of Method?

- Empirical Question
- Generally, if less than 5% of cases are missing data, method doesn’t matter.

Summary

- Is case deletion ever a good idea? No
- Is single imputation ever a good idea?
  - Yes, when little is to be gained from MI techniques and power would be lost in omitting techniques.
    - Less than 5% of cases have missing data.
- Do analysis multiple ways and see if you get different results. If yes, pick the most empirically justified method (EM or MI).

	Mean cholesterol at 14 days	Average decrease from day 2 to day 14	Correlation between days 2 and 14
EM Algorithm	222.2329	31.69567	0.4036006
DA Algorithm	221.7628	34.63067	0.3909713
Multiple Imputation	222.7931	31.13544	0.3871474
Complete cases	221.47368	38.00	0.392771
Mean imputation	221.47368	32.45489	0.3222948
Last-observation carried forward	224.28571	29.64286	0.470174



X1 X2……………………………Xp Y

1

…

…

…

…

…

….

..

N

Y_{obs}

Ways Data Can be Missing

- Missing at Random (MAR)
  - Missing values can depend on the value of observed variables (Xs) but not on the values of Y.
- Missing Not at Random (MNAR)
  - Missingness depends on value of Y even after relationships with Xs have been accounted for.
- Missing Completely at Random (MCAR)
  - Missing values are a random sample of all values. Missingness does not depend on Xs or Y

Examples

- Two Variable data set: Gender has been completely observed and aggression has missing data.
  - MAR: missingness might depend on gender but not level of aggression.
    - If more data is missing for men than women, but randomly within each of these distributions…ok
    - If level of aggression (beyond that predicted by gender) is predictive of missingness, cannot assume MAR.
  - MCAR: missingness is independent of gender and level of aggression.

How Realistic is MAR?

- Often need to assume it without being able to check (without extensive follow-up).
- Most recommended methods do well under moderate violations of MAR.

Resources

- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman & Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.
- Schimert, J., Schafer, J. L., Hesterberg, T., Fraley, C., & Clarkson, D. B. (2001). *Analyzing Data with Missing Values in S-Plus*. Seattle, WA: Insightful Corporation.
- http://www.multiple-imputation.com/

Software (http://www.multiple-imputation.com/)

- Missing Data Library in S-Plus 6. Based upon the work of Joseph L. Schafer, it features Gaussian, Loglinear and Conditional Gaussian. Performing multiple complete data analysis after multiple imputation, and consolidating results, is simplified by using the library.
- SOLAS for Missing Data Analysis 3.0 is a commercial Windows program by Statistical Solutions Limited. Version 3.0 offers new methods for multiple imputation, primarily based on Rubin's Chapter 5, pretty interface.
- NORM, CAT, MIX and PAN is software for multivariate imputed by Joseph L. Schafer. **NORM** uses a normal model. **CAT** uses a loglinear model for categorical data. **MIX** relies on the general location model for mixed categorical and continuous data. **PAN** is geared toward panel data. S-PLUS 3.3 and 4.0 and stand-alone Windows software is available.
- SPSS: As far as I can tell, for $400, you get missingness diagnostics and single imputations from EM.

Is HLM an Missing data Method?

- HLM is regression that handles non-independence of observations (clustering)
- Can analyses longitudinal data with different assessment schedules.
- Can analyses “unbalanced” clusters
- Keeps track of number and spacing of measurements.